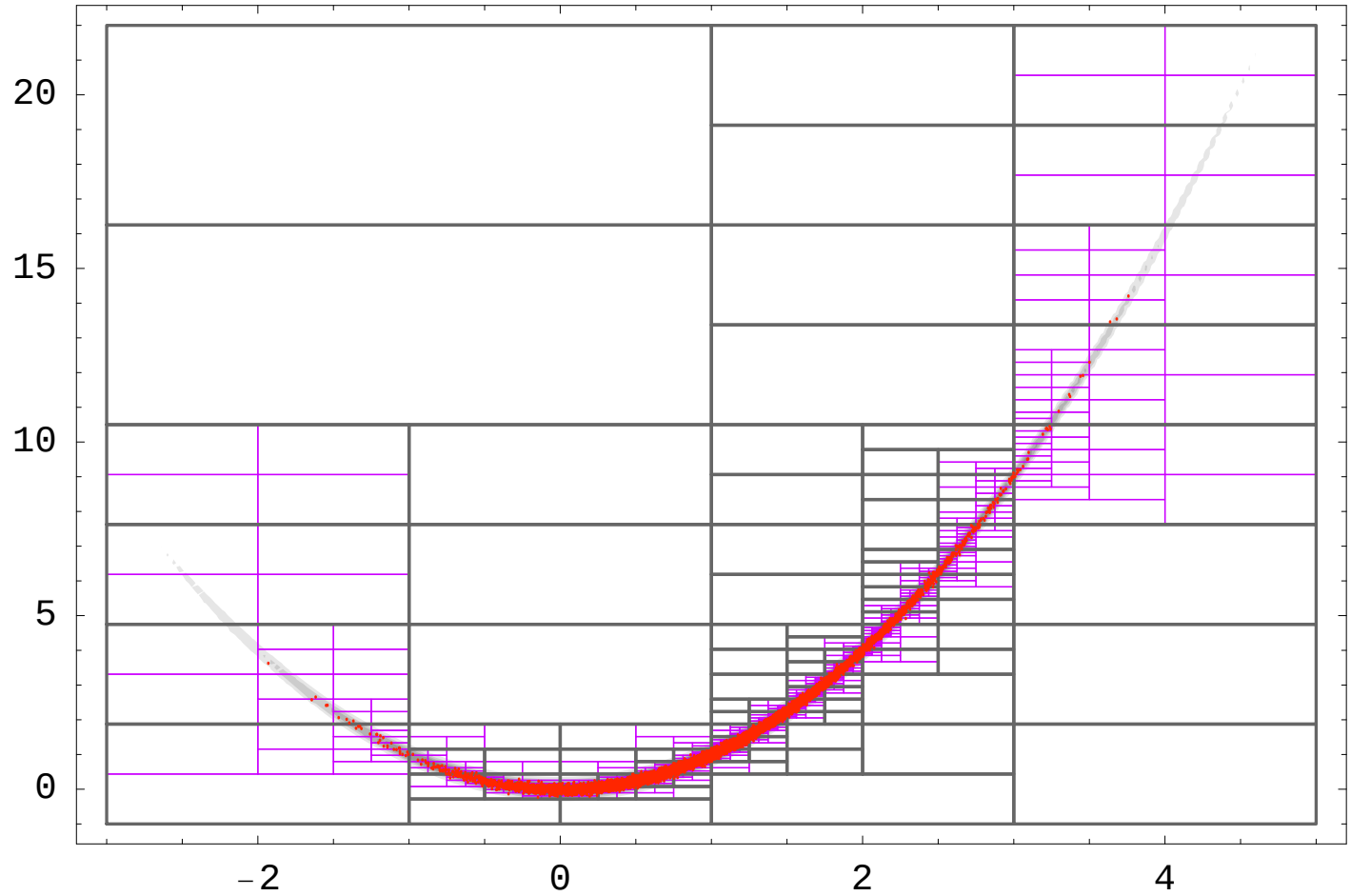
An auto-validating trans-dimensional von Neumann rejection sampler

Raazesh Sainudiin

Biomathematics Research Centre, Department of Mathematics & Statistics,

University of Canterbury, Christchurch, NZ

[with Tom York, Cornell University, Ithaca, NY, USA]



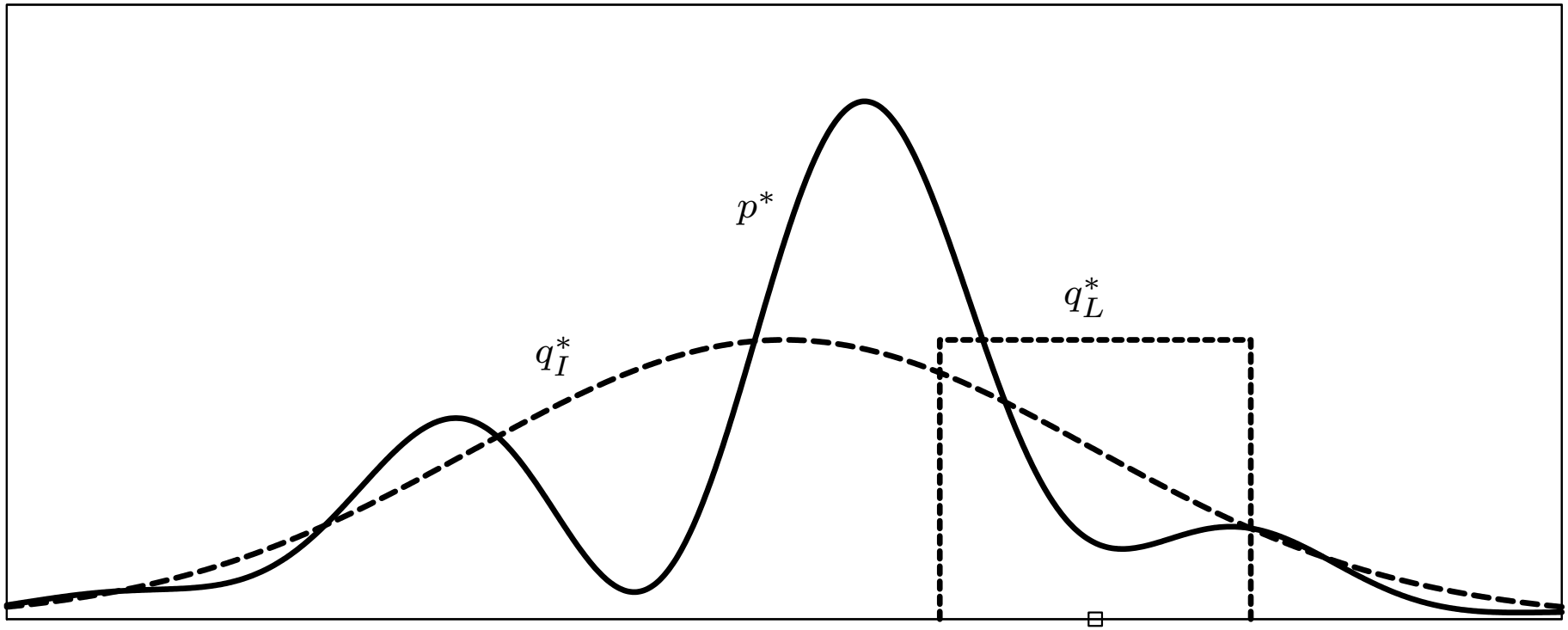
Outline

- A Statistical Sampling Problem
- Validated Numerics
  - What is it ?
  - How does it work ?
- A Solution to the Sampling Problem
  - Moore Rejection Sampling
  - Examples
- Summary and Extensions

Problem: Sampling From a Density

Sampling from a Density - MCMC (M-H)

Support : $\Theta \ni \theta$ Target : $p := p^*/N_p$ (Unknown N_p) Proposal : $q := q^*/N_q$



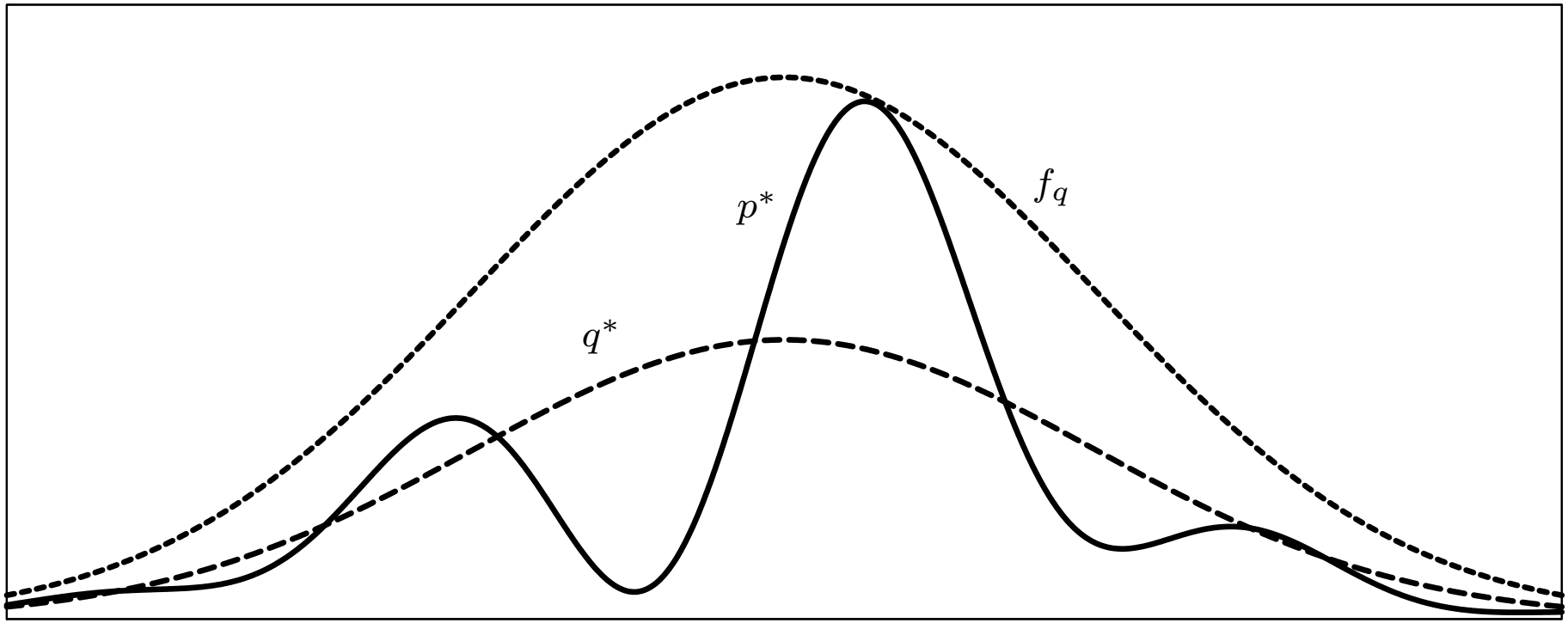
1. Choose an arbitrary starting point θ_0 and set $i = 0$.
2. Generate a candidate point $\theta' \sim q(\theta_i, \cdot)$ and $u \sim U(0, 1)$.
3. Set:

$$\theta_{i+1} = \begin{cases} \theta' & \text{if } u \leq \frac{p^*(\theta')q(\theta', \theta_i)}{p^*(\theta_i)q(\theta_i, \theta')} \\ \theta_i & \text{otherwise} \end{cases}$$

4. Set $i = i + 1$ and GO TO 2

Sampling from a Density - Rejection Sampling

Support : $\Theta \ni \theta$ Target : $p := p^*/N_p$ (Unknown N_p) Proposal : $q := q^*/N_q$



Find "envelope function" $f_q(\theta) = cq^*(\theta)$ such that $f_q(\theta) \geq p^*(\theta),\ \forall \theta \in \Theta$

1. Generate a candidate point $\theta \sim q(\cdot)$.
2. Draw $u \sim U(0,1)$.
3. If $(u < p^*(\theta)/f_q(\theta))$, then θ is an exact and independent sample from p. DONE.
4. Else GO TO 1

MCMC vs Rejection Sampling

MCMC (extensions of Metropolis-Hastings Chains)

- "Easy" to implement; almost any proposal works – BUT ONLY asymptotically...,
- Poor proposal $\Rightarrow$ slow convergence – heuristic proposal 'tuning',
- Convergence diagnostics are generally not rigorous – can be misleading.

Rejection Sampling (due to Von Neumann)

- "Hard" to implement; envelope property is NECESSARY – or will NOT sample p,
- Poor proposal $\Rightarrow$ low acceptance probability – MUCH $\ll 10^{-10}$,
- Perfectly independent samples – and NO convergence issues.

What makes a density hard to sample?

- Generally:
  1. Complexity – many peaks and valleys – size of Θ.
  2. Curse of dimensionality.
- If ignorant of global behavior of density:
  3. Widely separated peaks (hard to get from one to next),
  4. Narrow peaks on smooth background (hard to find),
  5. Peaks of strange shapes (e.g. Rosenbrock's banana density)
  6. Others... ?

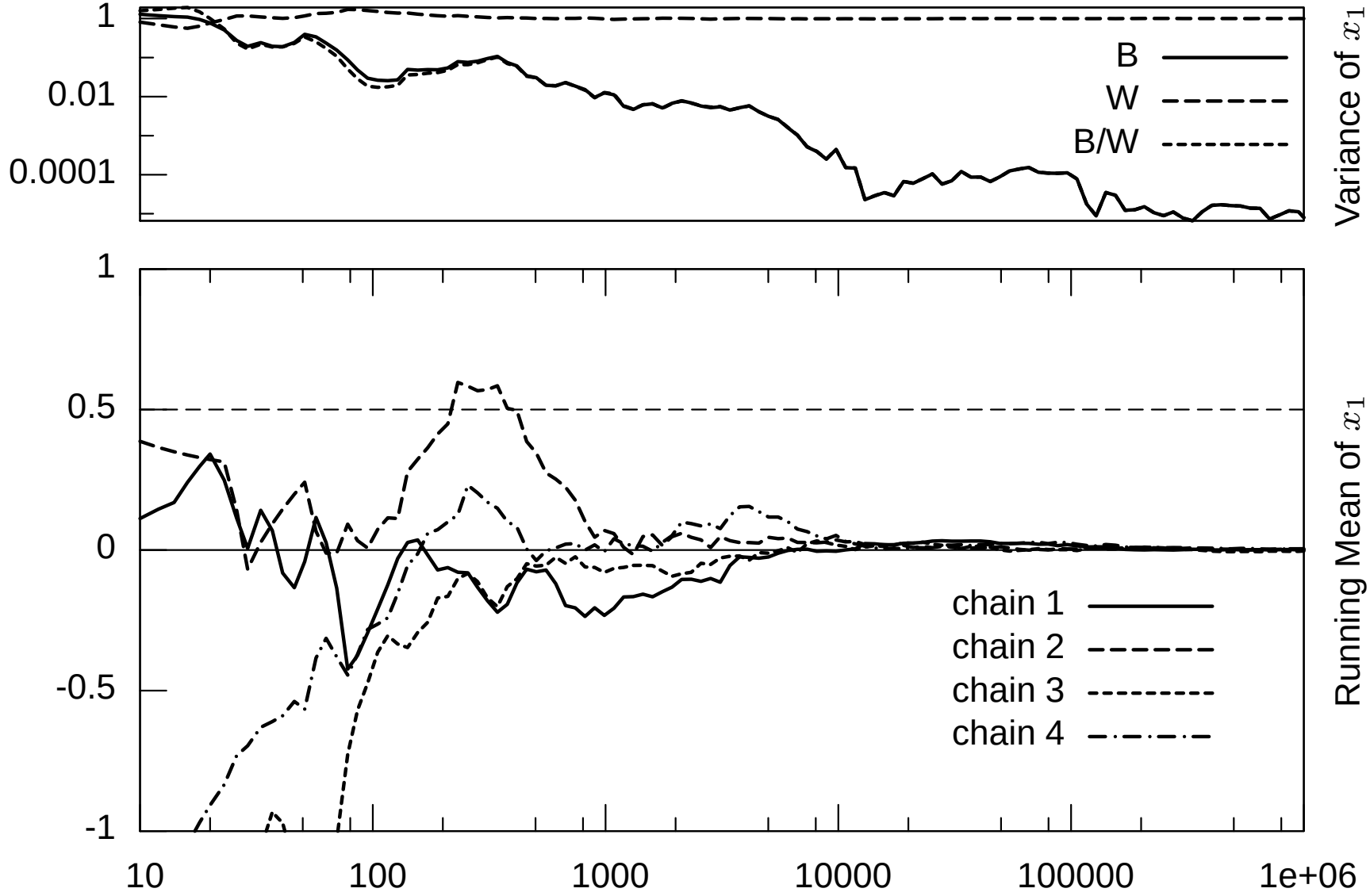
Don't treat target density as black box.

- Look at expression (or code) for p*(x).
- Find locations, widths, shapes of peaks, etc.
- Construct a better proposal.
- Interval arithmetic is way to do this in an auto-validated manner
  - AND we get envelope function.

Trivariate Needle in a Haystack – Heuristic Diagnostics

$$p^*(x) = \frac{1}{\sigma_1^3}\exp\{-\frac{1}{2}((x-\mu_1)/\sigma_1)^2\} + \frac{1}{\sigma_2^3}\exp\{-\frac{1}{2}((x-\mu_2)/\sigma_2)^2\}$$

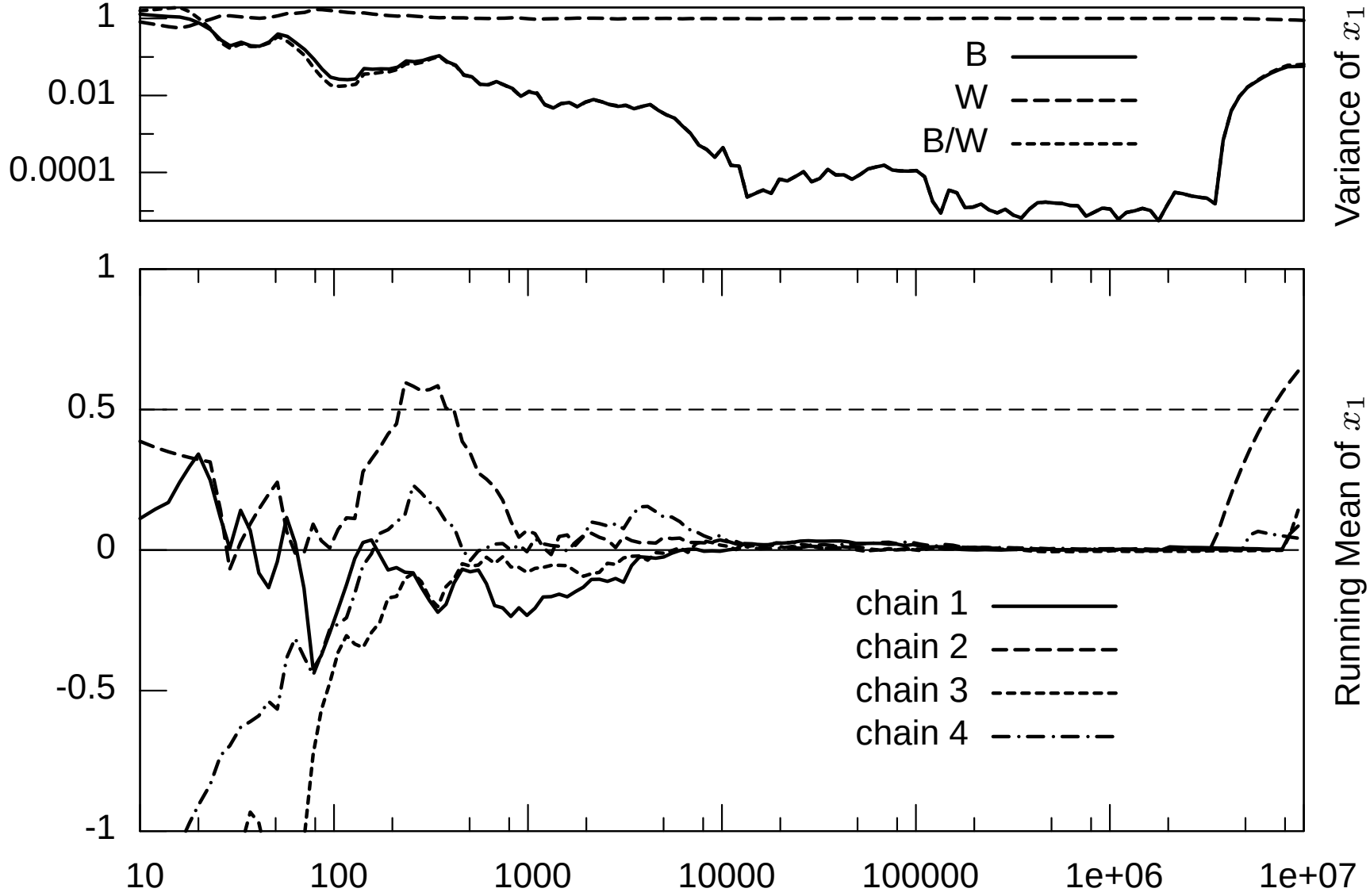
$$\mu_1 = (0,0,0), \mu_2 = (1,1,1), \sigma_1 = 1, \sigma_2 = 0.006$$



Trivariate Needle in a Haystack – Heuristic Diagnostics

$$p^*(x) = \frac{1}{\sigma_1^3}\exp\{-\frac{1}{2}((x-\mu_1)/\sigma_1)^2\} + \frac{1}{\sigma_2^3}\exp\{-\frac{1}{2}((x-\mu_2)/\sigma_2)^2\}$$

$$\mu_1 = (0,0,0), \mu_2 = (1,1,1), \sigma_1 = 1, \sigma_2 = 0.006$$



Some possibilities through

Validated Numerics ...

Validated Numerics

What is it ?

- set-valued mathematics
- intervals replace real numbers

Why use it ?

- provides rigorous error bounds
- naturally models uncertainty in data
- may produce faster numerical methods

Where has it been used recently ?

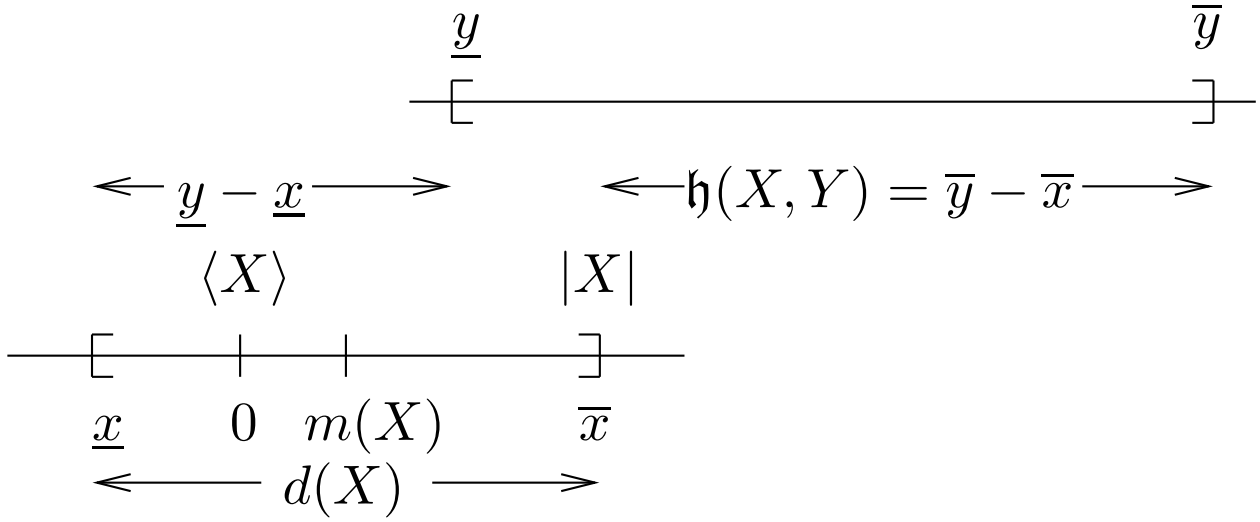
- J. Hass, M. Hutchings, and R. Schlafly, The Double Bubble Conjecture. Electr. Research Announcements of the Amer. Math. Soc., **1**, 98-102, 1995.
- W. Tucker, A Rigorous ODE Solver and Smale's 14th Problem, Found. Comput. Math. **2:1**, 53-117, 2002.
- T. C. Hales, Some algorithms arising in the proof of the Kepler conjecture. Discrete and computational geometry, **25**, 489-507, Algorithms Combin., 2003.

Early work:

R. C. Young (1931), M. Warmus (1956), T. Sunaga (1958), R. E. Moore (1959, 1966).

Intervals

Notations, Definitions and Features



- real: $x \in \mathbb{R}$
- (compact) real interval: $X = [\underline{x}, \overline{x}] = [\inf(X), \sup(X)]$
- set of all real intervals: $\mathbb{IR} := \{[a,b] : a \leq b,\ a,b \in \mathbb{R}\}$
  - Example: $[1,\pi], 17, \sqrt{2} \in \mathbb{IR}$, but not $[2,1]$ or $[1,\infty]$.
- real interval vector or box: $X = (X_1, \cdots, X_n)^T \in \mathbb{IR}^n$, where $X_i = [\underline{x_i}, \overline{x_i}] \in \mathbb{IR}$, $1 \leq i \leq n$
- a thin interval $X = [x,x]$ has 0 diameter with $\underline{x} = \overline{x} = x \Rightarrow \mathbb{R} \subset \mathbb{IR}$

Arithmetic Over $\mathbb{IR}$

Definition. *If* $\circ$ *is one of the operators* $+, -, /, \cdot$ *and if* $X, Y \in \mathbb{IR}$

$$X \circ Y := \{x \circ y : x \in X, y \in Y\}$$

except that X/Y *is undefined if* $0 \in Y$.

Uncountable many cases to consider!

Continuity, Monotonicity, and Compactness $\Rightarrow$

$$X + Y = [\underline{x} + \underline{y}, \overline{x} + \overline{y}], \qquad X \cdot Y = [\min\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\}, \max\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\}],$$

$$X - Y = [\underline{x} - \overline{y}, \overline{x} - \underline{y}], \text{ and} \qquad X/Y = X \cdot [1/\overline{y}, 1/\underline{y}],\ 0 \notin Y.$$

On a computer we use directed rounding:

$$X + Y = [\bigtriangledown(\underline{x} \oplus \underline{y}), \bigtriangleup(\overline{x} \oplus \overline{y})]$$

We then have $X \circ Y \supseteq \{x \circ y : x \in X, y \in Y\}$

Interval Extensions

One of the the main goals is to enclose the range of a real-valued function f :

$$R(f; D) := \{f(x) : x \in D\}$$

This is achieved by constructing an interval extension $F : \mathbb{IR} \to \mathbb{IR}$ of $f : \mathbb{R} \to \mathbb{R}$

Monotone functions are easy !

$$\begin{aligned}
\exp(X) &= [\exp(\underline{x}), \exp(\overline{x})] \\
\sqrt{(}X) &= [\sqrt{(}\underline{x}), \sqrt{(}\overline{x})], && \text{if } 0 \leq \underline{x} \\
\log(X) &= [\log(\underline{x}), \log(\overline{x})], && \text{if } 0 < \underline{x} \\
\arctan(X) &= [\arctan(\underline{x}), \arctan(\overline{x})].
\end{aligned}$$

Piecewise monotone functions are also OK!

$$X^n = \begin{cases} [\underline{x}^n, \overline{x}^n] & : \text{if } n \in \mathbb{Z}^+ \text{ is odd,} \\ [\langle X \rangle^n, |X|^n] & : \text{if } n \in \mathbb{Z}^+ \text{ is even,} \\ [1, 1] & : \text{if } n = 0, \\ [1/\overline{x}, 1/\underline{x}]^{-n} & : \text{if } n \in \mathbb{Z}^-; 0 \notin X \end{cases}$$

Class of Standard Elementary Functions $\mathfrak{E}$

We define the class of standard functions to be the set

$$\mathfrak{S} = \{\exp x, \log x, x^a, |x|, \sin x, \cos x, \tan x, \cdots, \arccos x, \arctan x, \sinh x, \cosh x, \tanh x\}$$

For any $f \in \mathfrak{S}$, we can construct a *sharp* interval extension F, i.e.,

$$f \in \mathfrak{S} \Rightarrow R(f; X) = F(X).$$

Building new functions is easy...

We use finite combinations of constants, elements of $\mathfrak{S}$, $\{+, -, \cdot, /\}$, and their compositions to build the elementary functions $\mathfrak{E}$. Interval versions of $\mathfrak{S}$ and $\{+, -, \cdot, /\}$ provide the corresponding interval extensions.

But, we may now over-estimate the range ...

If $f(x) = \frac{x}{1+x^2}$, then $F(X) = \frac{X}{1+X^2}$. For the interval $X = [1, 2]$, we have

$$R(f; [1, 2]) = [\frac{2}{5}, \frac{1}{2}] \subseteq [\frac{1}{5}, 1] = F([1, 2]).$$

Interval Enclosures

Theorem (1). *If $f(x) \in \mathfrak{E}$, and $F(X)$ is well-defined, then*

$$R(f;X) \subseteq F(X)$$

How tight is the enclosure ?

Theorem (2). *If $f \in \mathfrak{E}$, $X = X_1 \cup X_2 \cdots \cup X_k$, and $F(X)$ is well-defined, then*

$$R(f;X) \subseteq \bigcup_{i=1}^{k} F(X_i) \subseteq F(X)$$

If f is Lipschitz on X there is a $K \geq 0$ s.t.

$$d\left(\bigcup_{i=1}^{k} F(X_i)\right) - d\left(R(f;X)\right) = K \max_i d(X_i)$$

I.A. (almost) gives us access to $R(f;X)$.

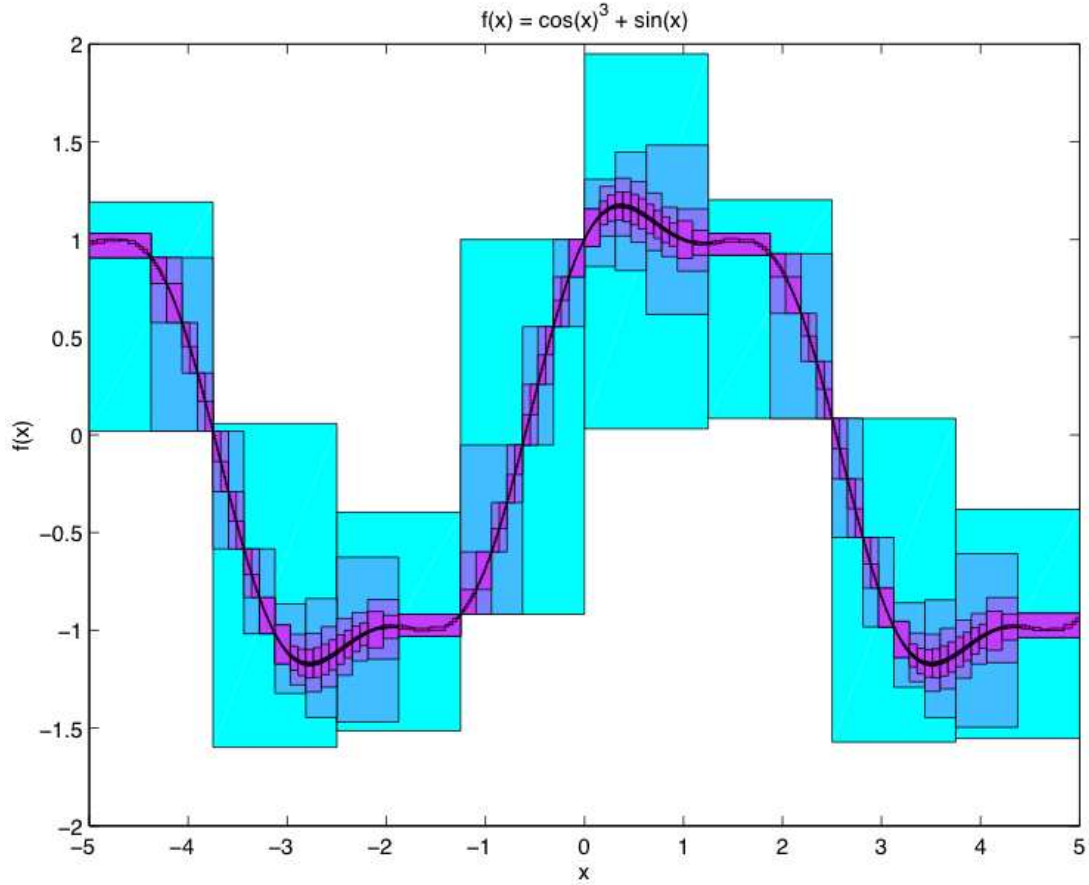
Computer-Aided Proofs

A consequence of Theorem 1 : $y \notin F(X) \Rightarrow y \notin R(f;X)$

Exercise 1: Let $f(x) = \cos(x)^3 + \sin(x)$. Prove that $f(x) \neq 0$, $\forall x \in [0, \frac{5}{4}]$.

Solution: Define $F(X) = \cos(X)^3 + \sin(X)$. Then, by Theorem 1, we have

$$R(f;[0,\frac{5}{4}]) \subseteq F([0,\frac{5}{4}]) = [\cos(\frac{5}{4})^3, 1] + [0, \sin(\frac{5}{4})] \subseteq [0.0313, 1.9490].$$

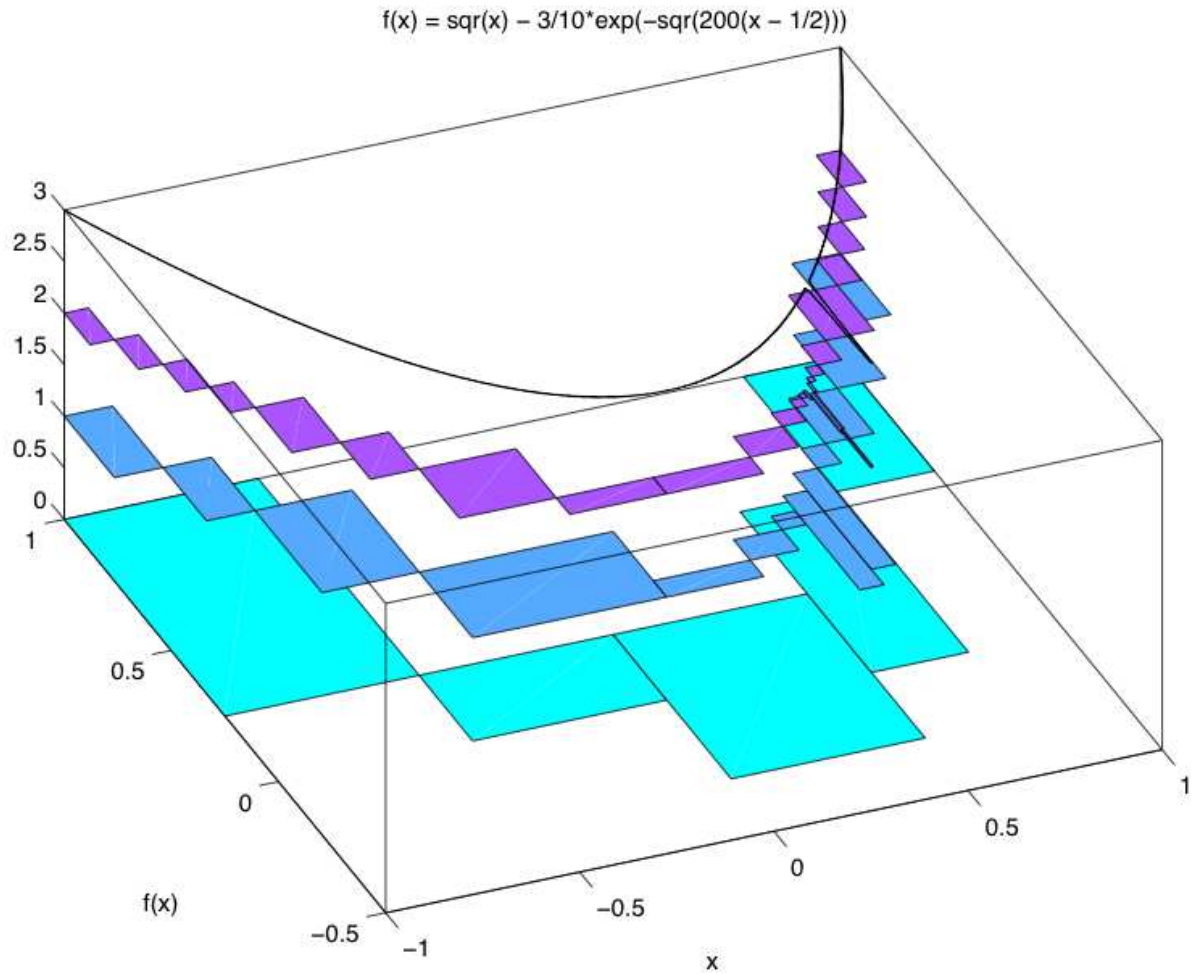


'Impossible' cases too

Exercise 2: Draw the graph of the function

$$f_a(x) = x^2 - \frac{3}{10} e^{-(a(x-\frac{1}{2}))^2}, \text{ for } a = 200, \text{ over the interval } [-1, 1].$$

Even for huge a, the I.A.-methods *cannot* miss the sharp bend! Conventional methods do.



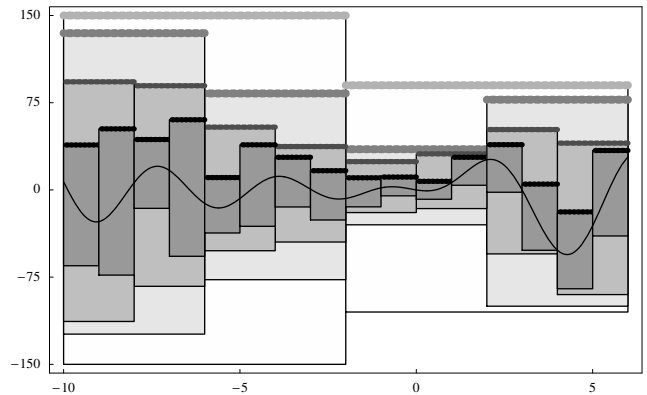
Rejection Envelopes via Computer-Aided Proofs

Obtain an envelope of the Lipschitz function $-\sum_{k=1}^{5} k\,x\,\sin\left(\frac{k(x-3)}{3}\right)$.

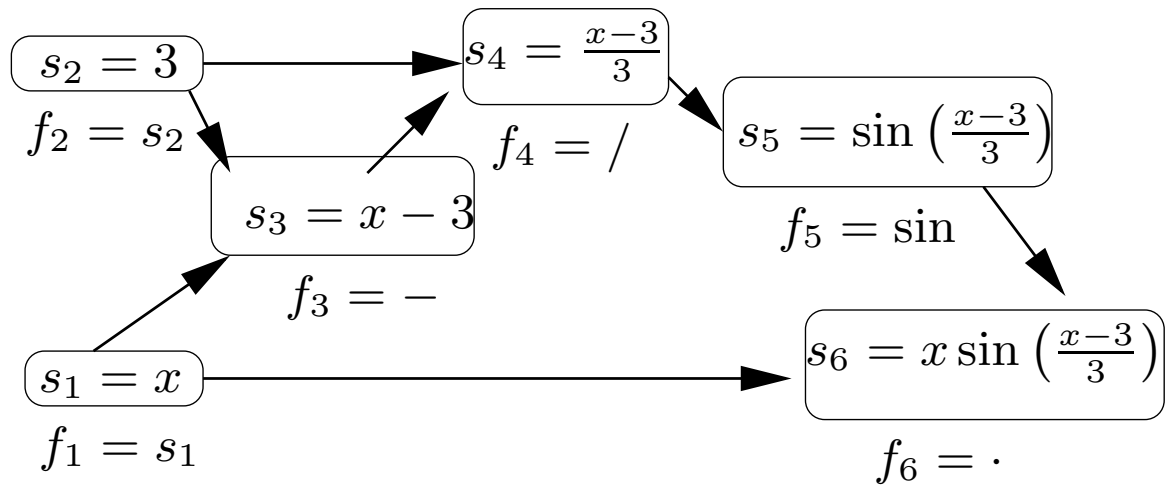
Solution: Define $F(X) = -\sum_{k=1}^{5} k\,X\,\sin\left(\frac{k(X-3)}{3}\right)$. Then, by

Theorem 1 $R(f;X) \subseteq F(X)$

Theorem 2 we can bisect the domain into smaller pieces until $\max_i d\left(F(X_i)\right) \leq \texttt{TOL}$



Recursive evaluation of the sub-expressions s_i by f_i on the DAG for $x \cdot \sin((x-3)/3)$



Auto-validating von Neumann RS ⇔ Moore RS (MRS)

Suppose,

Compact domain	$\mathbf{\Theta} = [\underline{\theta}, \overline{\theta}]$
Target shape	$p^*(\theta) : \mathbf{\Theta} \to \mathbb{R}$
Target integral	$N_p := \int_{\mathbf{\Theta}} p^*(\theta)\, d\theta$
Target density	$p(\theta) := \frac{p^*(\theta)}{N_p} : \mathbf{\Theta} \to \mathbb{R}$
Interval extension of p^*	$P^*(\Theta) : \mathbb{I}\mathbf{\Theta} \to \mathbb{IR}$
Partition of$\mathbf{\Theta}$	$\mathfrak{T} := \{ \Theta^{(1)}, \Theta^{(2)}, ..., \Theta^{(\lvert\mathfrak{T}\rvert)} \}$

then, by Theorem 1

$$p^*(\Theta^{(i)}) \subseteq P^*(\Theta^{(i)}) := [\underline{P}^*(\Theta^{(i)}), \overline{P}^*(\Theta^{(i)})],\ \forall\, i \in \{1, 2, ..., |\mathfrak{T}|\}.$$

Construct the $\mathfrak{T}$-specific proposal $q^{\mathfrak{T}}(\theta)$ as a normalized simple function over $\mathbf{\Theta}$

$$q^{\mathfrak{T}}(\theta) = \left(N_{q^{\mathfrak{T}}}\right)^{-1} \sum_{i=1}^{|\mathfrak{T}|} \overline{P}^*(\Theta^{(i)})\, \mathbf{1}_{\{\theta\, \in\, \Theta^{(i)}\}}, \qquad N_{q^{\mathfrak{T}}} := \sum_{i=1}^{|\mathfrak{T}|} \left(d(\Theta^{(i)}) \cdot \overline{P}^*(\Theta^{(i)})\right).$$

Then an envelope function $f(q^{\mathfrak{T}}(\theta))$ guaranteeing the necessary inequality is

$$f_{q^{\mathfrak{T}}}(\theta) = \sum_{i=1}^{|\mathfrak{T}|} \overline{P}^*(\Theta^{(i)})\, \mathbf{1}_{\{\theta\, \in\, \Theta^{(i)}\}} \geq p^*(\theta),\ \forall\, \theta \in \mathbf{\Theta}$$

Auto-validating von Neumann RS ⇔ Moore RS (MRS)

Suppose,

Compact domain	$\mathbf{\Theta} = [\underline{\theta}, \overline{\theta}]$
Target shape	$p^*(\theta) : \mathbf{\Theta} \to \mathbb{R}$
Target integral	$N_p := \int_{\mathbf{\Theta}} p^*(\theta)\, d\theta$
Target density	$p(\theta) := \frac{p^*(\theta)}{N_p} : \mathbf{\Theta} \to \mathbb{R}$
Interval extension of p^*	$P^*(\Theta) : \mathbb{I}\mathbf{\Theta} \to \mathbb{IR}$
Partition of$\mathbf{\Theta}$	$\mathfrak{T} := \{ \Theta^{(1)}, \Theta^{(2)}, ..., \Theta^{(\lvert\mathfrak{T}\rvert)} \}$

Efficiency ⇔ Large average acceptance probability

$$\mathbf{A}_{\mathfrak{T}}^{p} = \frac{\int_{\mathbf{\Theta}} p^*(\theta)\, \partial\theta}{\int_{\mathbf{\Theta}} f_q(\theta)\, \partial\theta} = \frac{N_p}{\sum_{i=1}^{|\mathfrak{T}|} \left(d(\Theta^{(i)}) \cdot \overline{P}^*(\Theta^{(i)}) \right)} \geq \frac{\sum_{i=1}^{|\mathfrak{T}|} \left(d(\Theta^{(i)}) \cdot \underline{P}^*(\Theta^{(i)}) \right)}{\sum_{i=1}^{|\mathfrak{T}|} \left(d(\Theta^{(i)}) \cdot \overline{P}^*(\Theta^{(i)}) \right)}$$

Furthermore, if $p^* \in \mathfrak{E}_{\mathfrak{L}}$, the Lipschitz class of elementary functions

$$\mathbf{A}_{\mathfrak{T}}^{p} \nearrow 1 - \mathcal{O}\left(\max_{i \in \{1,\ldots,\mathfrak{T}\}} d(\Theta^{(i)}) \right).$$

Efficiency can be further improved by:

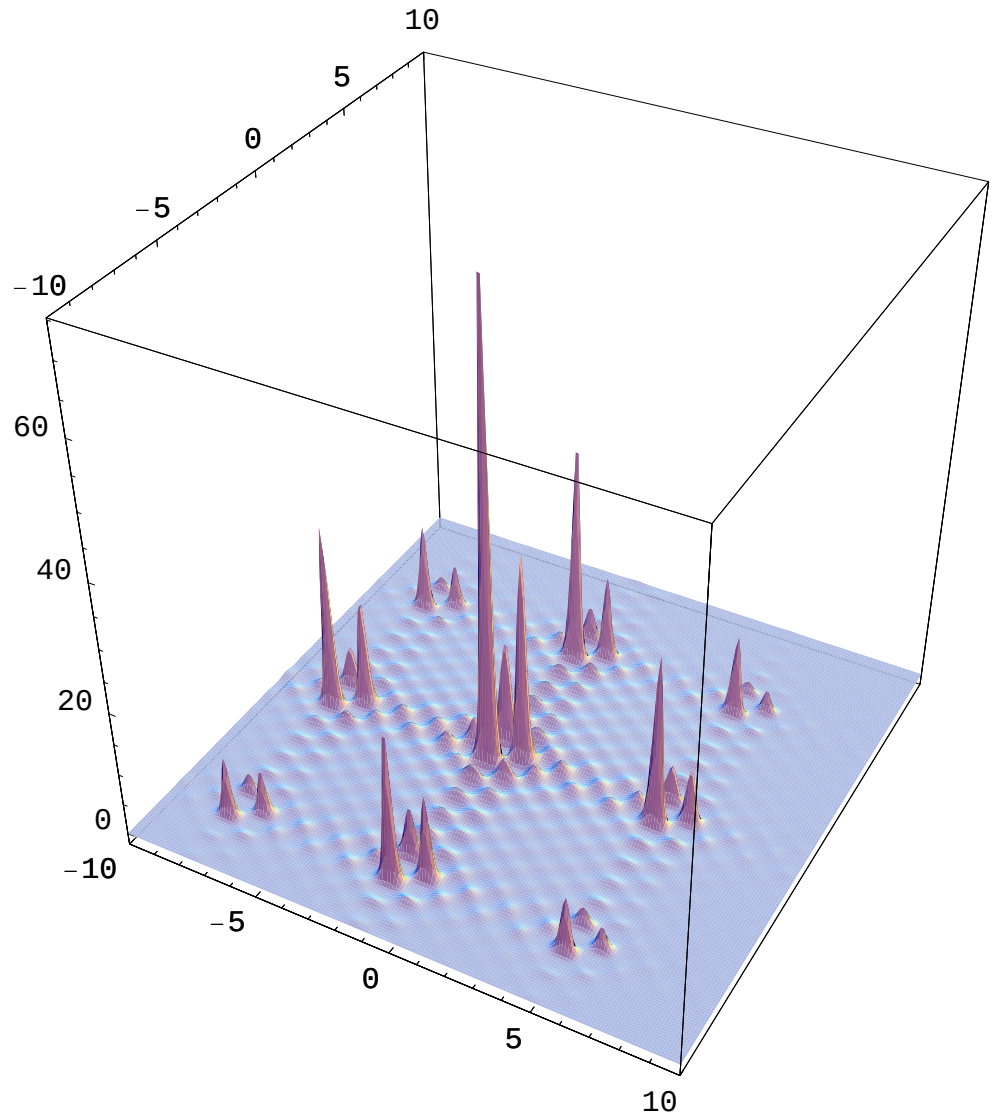
- tighter range enclosures through automatic differentiation
- clever partitioning strategies

MRS – Bivariate Levy Densities

$$E(X_1, X_2) = \sum_{i=1}^{5} i \cos\left((i-1)X_1 + i\right) \sum_{j=1}^{5} j \cos\left((j+1)X_2 + j\right) + (X_1 + 1.42513)^2 + (X_2 + 0.80032)^2$$

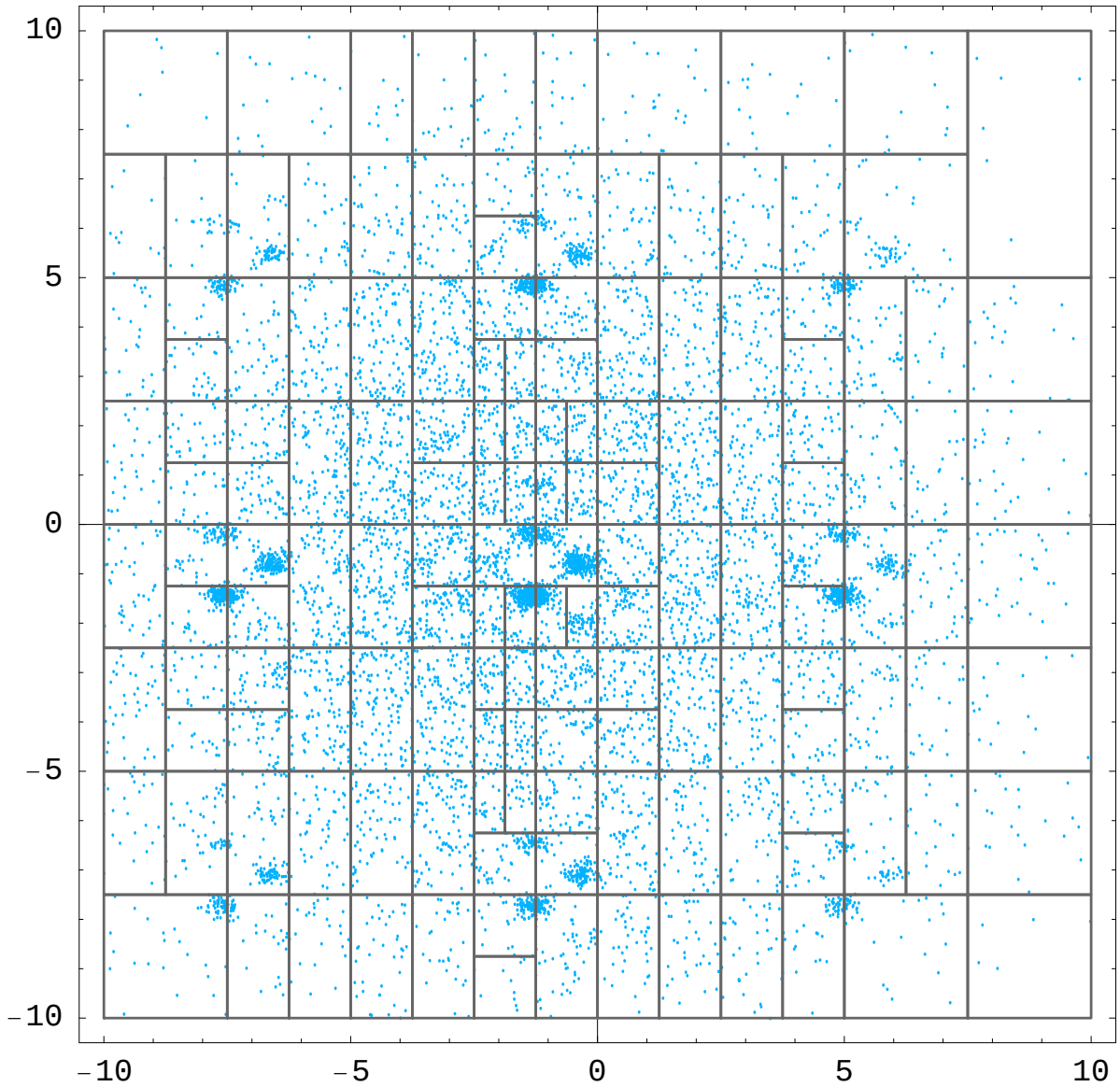
$$l_T(X_1, X_2) = \exp\{-E(X_1, X_2)/T\}$$

There are 700 modes !



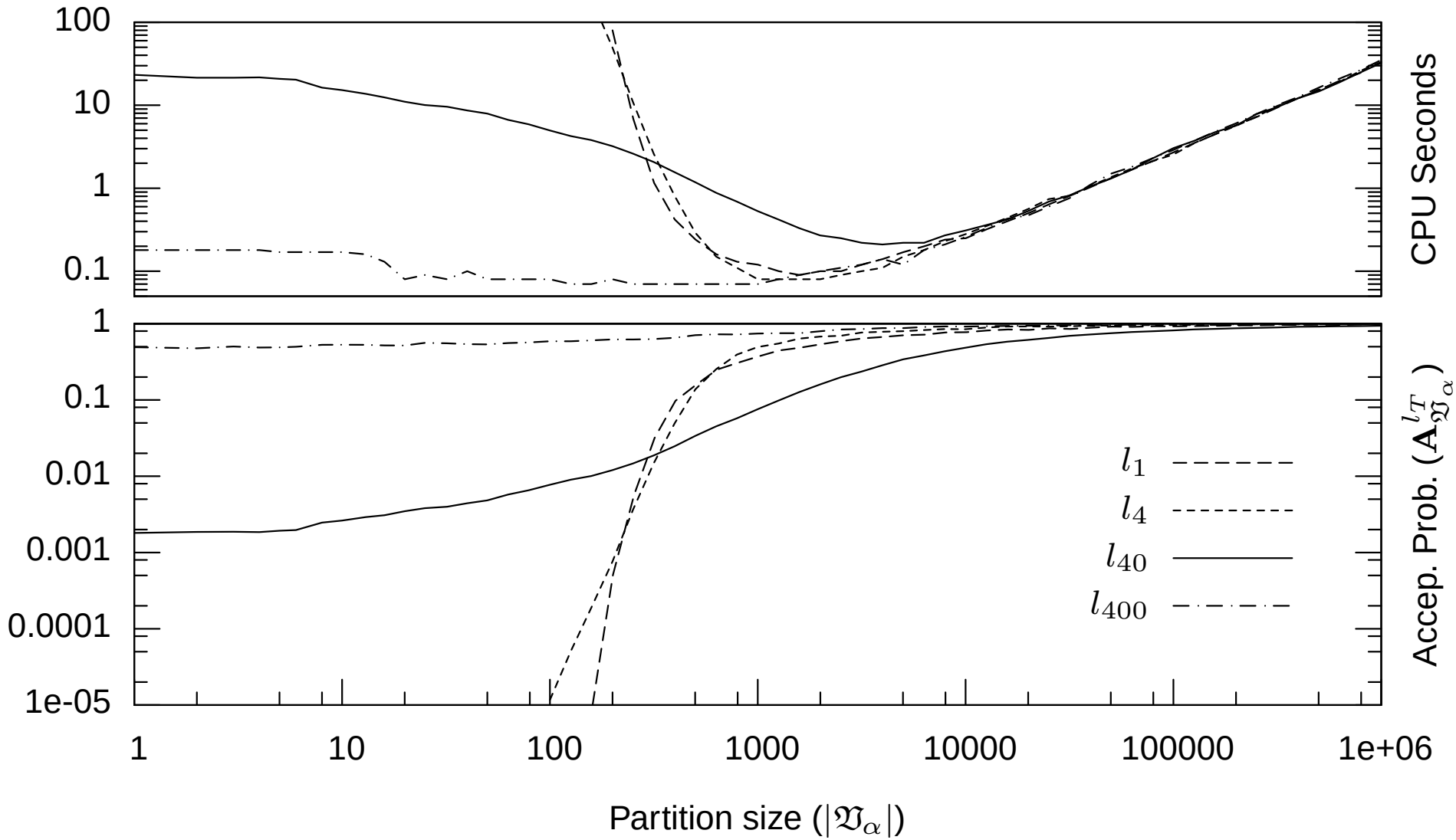
MRS – Bivariate Levy Densities

Adaptive partitioning of the domain $[-10, 10] \times [-10, 10]$ into 150 rectangles for Moore rejection sampling from the Levy target density l_{40} (acceptance probab. $= 0.01$).



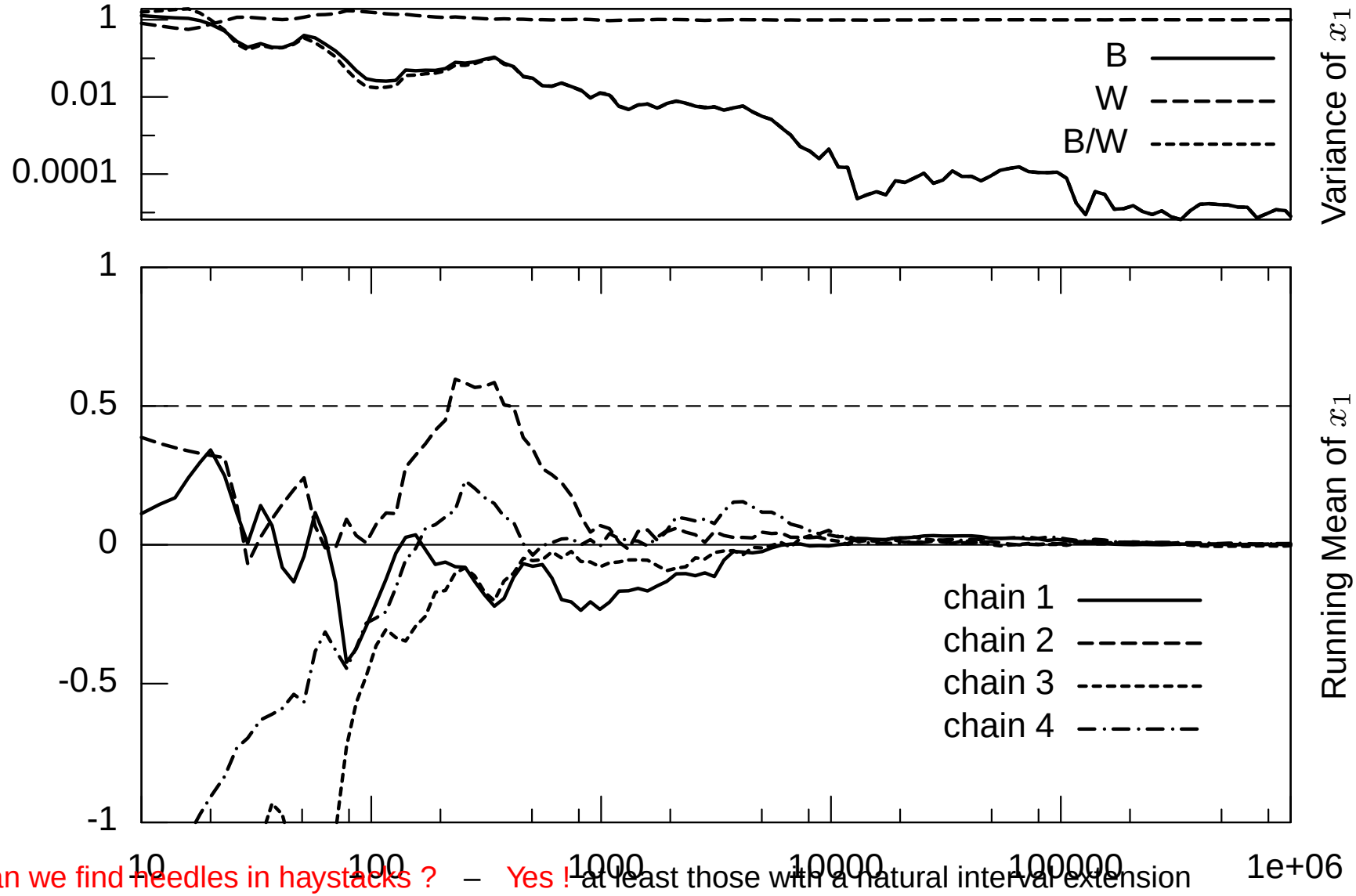
MRS – Bivariate Levy Densities

Acceptance probability ($\mathbf{A}_{\mathfrak{V}_\alpha}^{l_T}$) versus partition size ($|\mathfrak{V}_\alpha|$) for Levy targets l_T, where T is the is the temperature parameter. There is an optimal CPU time (2.0GHz) to generate 10^4 samples.



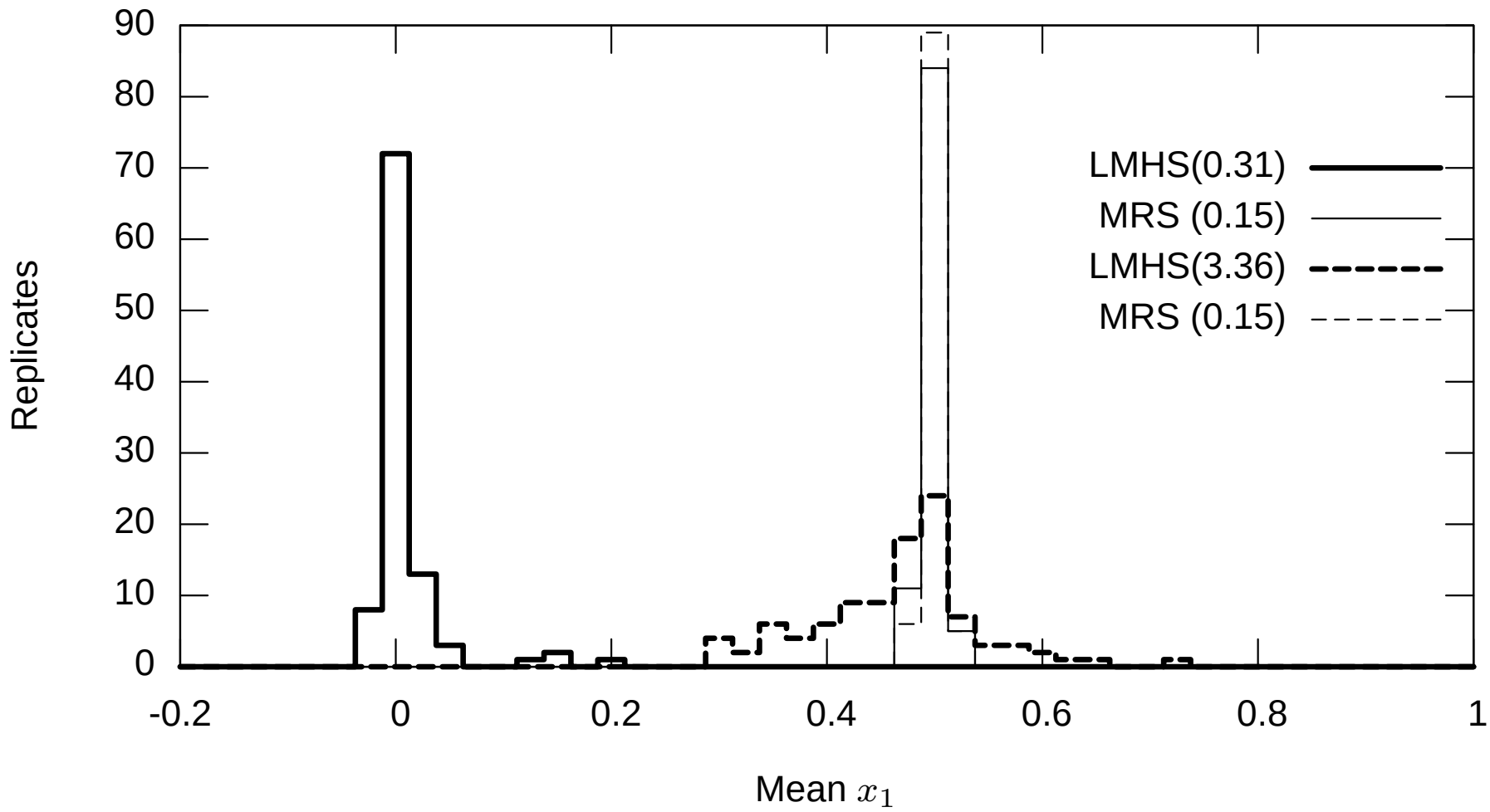
MRS – Trivariate Needle in a Haystack

$$p^*(x) = \frac{1}{\sigma_1^3}\exp\{-\frac{1}{2}((x-\mu_1)/\sigma_1)^2\} + \frac{1}{\sigma_2^3}\exp\{-\frac{1}{2}((x-\mu_2)/\sigma_2)^2\}$$



Can we find needles in haystacks ? – Yes ! – at least those with a natural interval extension

MRS – Trivariate Needle in a Haystack

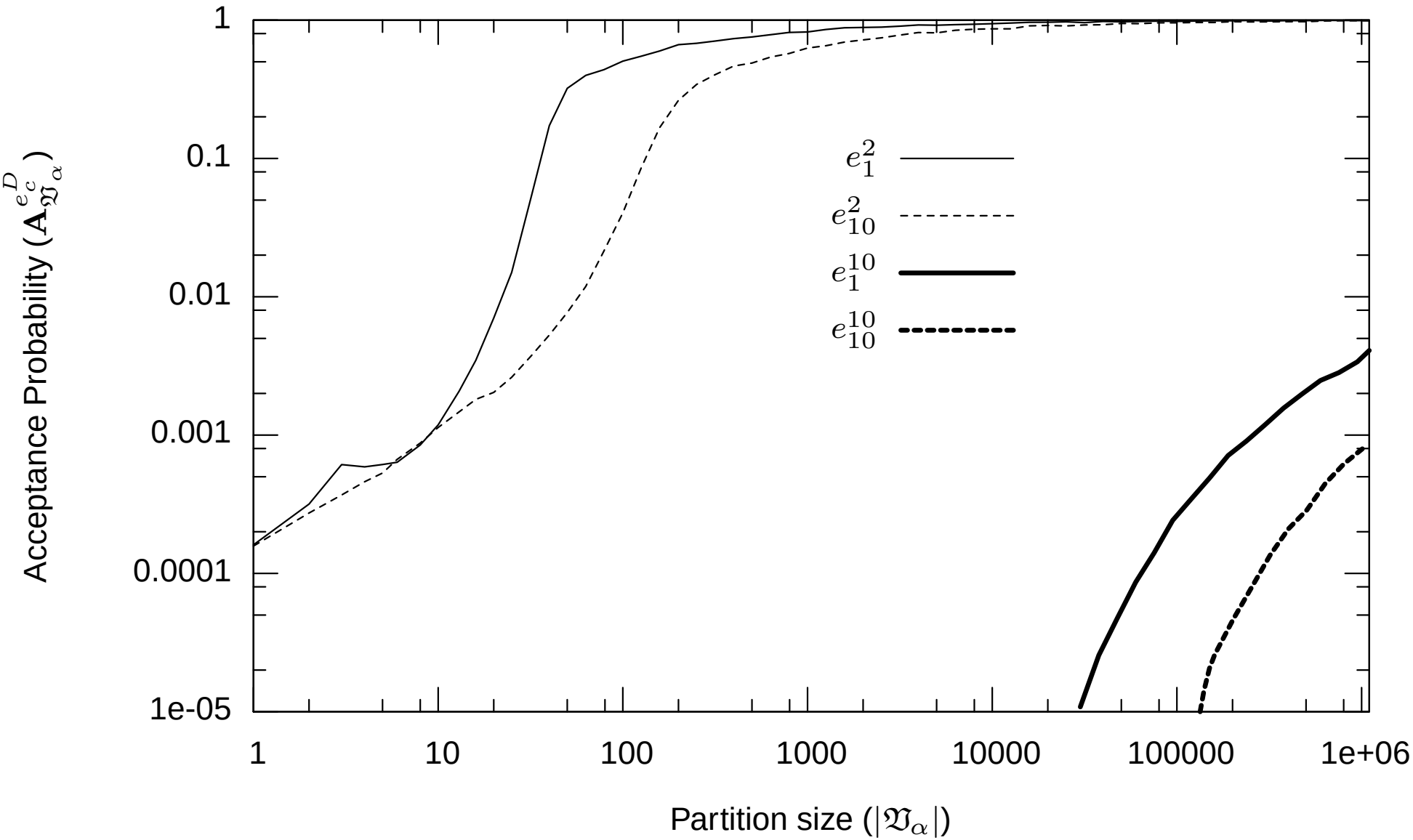


1
- MCMC with Metropolis proposal (uniform in cube of side $6\sigma_1$ centered at x)
- With burn-in defined as ending at $B/W = 0.05$
- Run length is 10 times burn-in (typical run length 20000-50000 for $\sigma_2 = 0.01$).

2 MRS with 1000 boxes and 10000 samples.

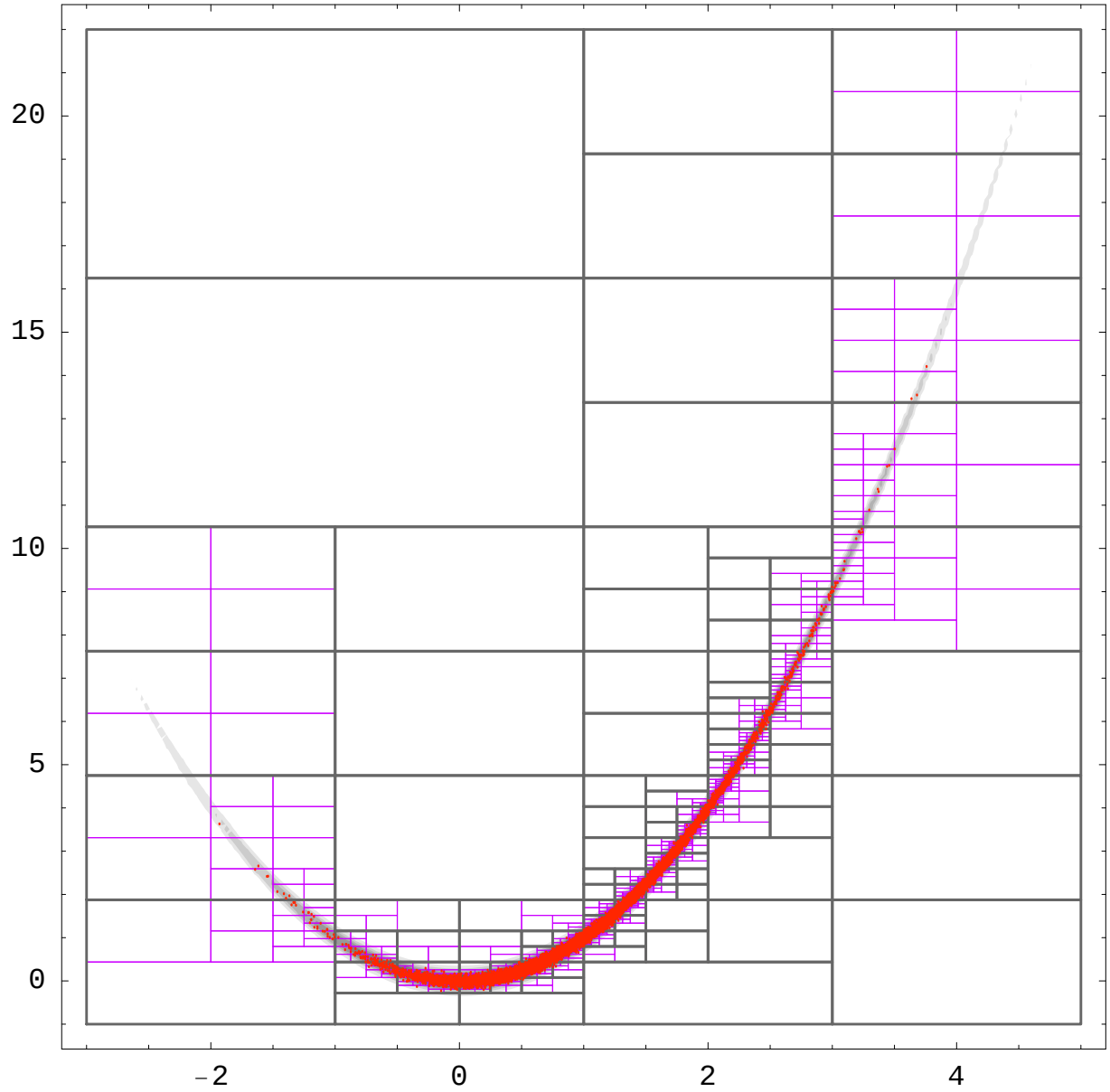
MRS – Multivariate Equi-Exponential Mixtures

$$e_c^D(X) = \sum_{i=1}^{c} \exp\left(-|X - \alpha^{(i)}|\right), \quad \alpha_j^{(i)} = a^{(i)} \in \Theta = [-100, 100]^D,\ j = 1, \ldots, D,\ i = 1, \ldots, c$$



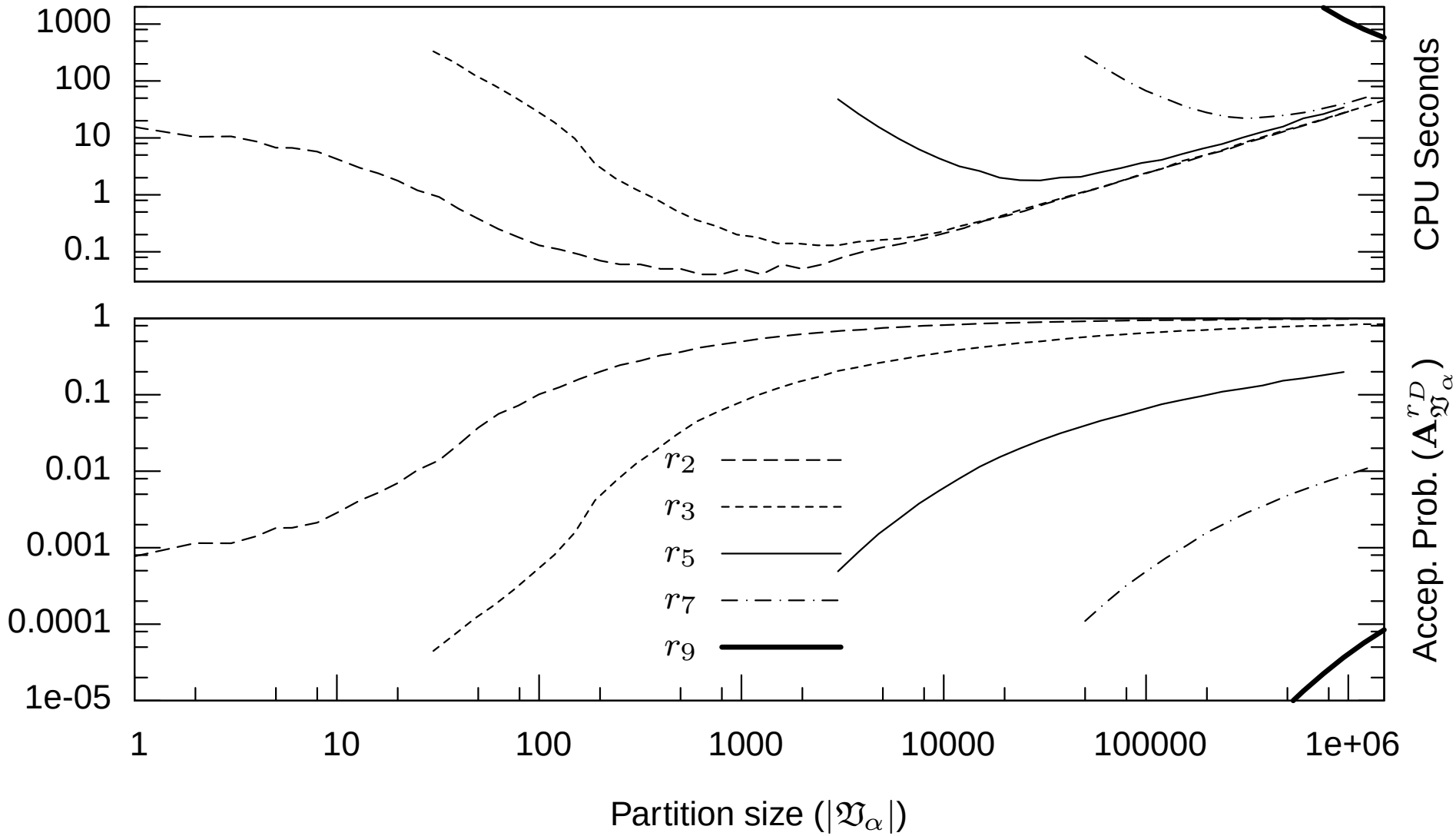
MRS – Multivariate Rosenbrock's Density

$$r_D(X) = \exp\{-\sum_{i=2}^{D}(100(X_i - X_{i-1}^2)^2 + (1 - x_{i-1})^2)\}$$



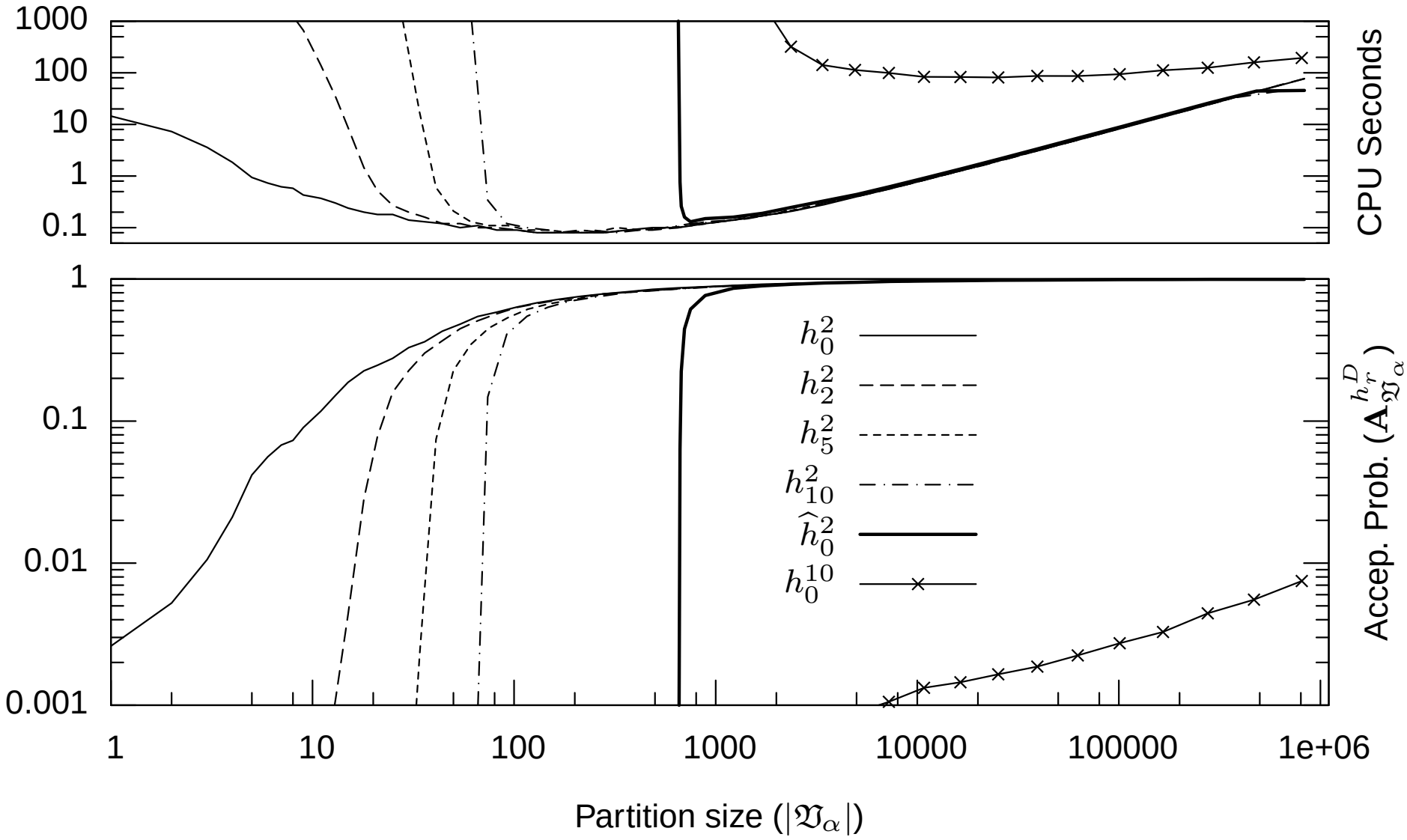
MRS – Multivariate Rosenbrock's Density

Acceptance probability ($\mathbf{A}^{r_D}_{\mathfrak{V}_\alpha}$) versus partition size ($|\mathfrak{V}_\alpha|$) for Rosenbrock targets r_D, where D is the is the dimension ($\boldsymbol{\Theta} = [-10, 10]^D$). CPU time (2.0GHz) to generate 10^4 samples.



MRS – Multivariate Witch's Hats

Acceptance probability ($\mathbf{A}_{\mathfrak{V}_\alpha}^{h_r^D}$) versus partition size ($|\mathfrak{V}_\alpha|$) for Witch's Hat targets h_r^D, where D is the support dimension and $R = 10^{-r}$ is the hat's radius.



Phylogenetic Likelihood

1: *input:* (i) a tree $\mathcal{T}_k$, (ii) branch lengths $t = (t_1, t_2, \ldots, t_{b_k})$, (iii) transition probability $P_{a_i,a_j}(t)$ for any $a_i, a_j \in \mathfrak{A}$, (iv) stationary distribution $\pi(a_i)$ over each characters $a_i \in \mathfrak{A}$, (v) site pattern (data) at site q

2: *output:* $\ell_q(k,t)$, the likelihood at site q

3: *initialize:* For a leaf node h with observed character a_i at site q, set $l_h^{(a_i)} = 1$ and $l_h^{(a_j)} = 0$ for all $j \neq i$. For any internal node h, set $l_h := (1, 1, \ldots, 1)$.

4: *recurse:* compute l_h for each sub-terminal node h, then those of their ancestors recursively to finally compute l_r for the root node r to obtain the likelihood for site q,

$$\ell_q(k,t) = l_r = \sum_{a_i \in \mathfrak{A}} \left(\pi(a_i) \cdot l_r^{(a_i)}\right).$$

For an internal node h with descendants $s_1, s_2, \ldots, s_\hbar$,

$$l_h^{(a_i)} = \sum_{j_1,\ldots,j_\hbar \in \mathfrak{A}} \{\, l_{s_1}^{(j_1)} \cdot P_{a_i,j_1}(t_{s_1}) \cdot l_{s_2}^{(j_2)} \cdot P_{a_i,j_2}(t_{s_2}) \ldots l_{s_\hbar}^{(j_\hbar)} \cdot P_{a_i,j_\hbar}(t_{s_\hbar}) \,\}.$$

Auto-validating Posterior Estimates of Human-Neandertal Divergence Time

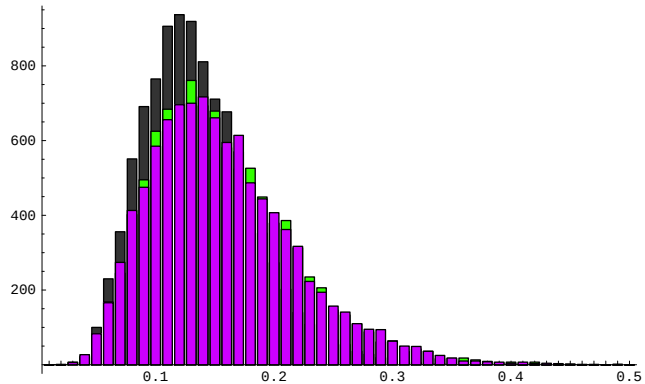
Envelope via Interval-extended post-order traversals

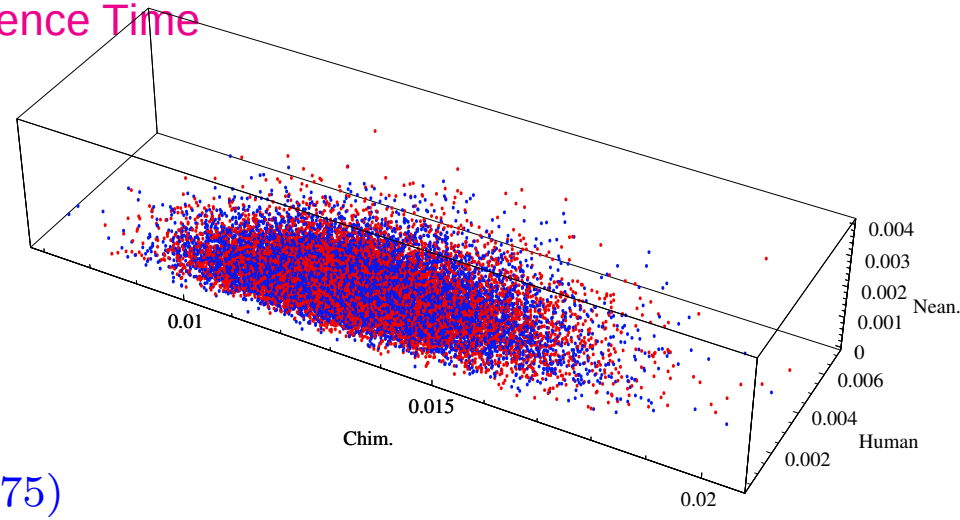
```
site       :                     1 1 1 1 1 1
pattern    : 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
 . . . . . . . . . . . . . . . . . . . . . .
neandertal : t t c a g g t g t c a a c a a
human      : t t c a g g t a c c a g t a g
chimpanzee : t c c a g a a a t t g a c t g
 . . . . . . . . . . . . . . . . . . . . . .
site       : 6 1 6 6 4 1 2 1 2 1 1 1 1 1 1
pattern    : 0 4 0 8 5 0       4 5
counts     : 5   3 5 0
```

Human-Neandertal Divergence Estimate $(461000, 821000)$ of Green et. al. (Nature, 2006) is too narrow

$10,000$ i.i.d. samples from the posterior over the Chimpanzee, Human and Neandertal phylogenies

THE Pos. Distn. of Human-Neandertal Divergence Time

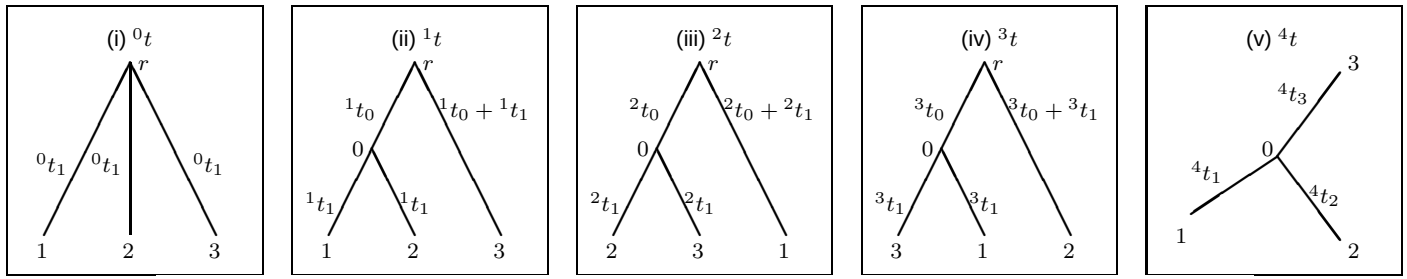


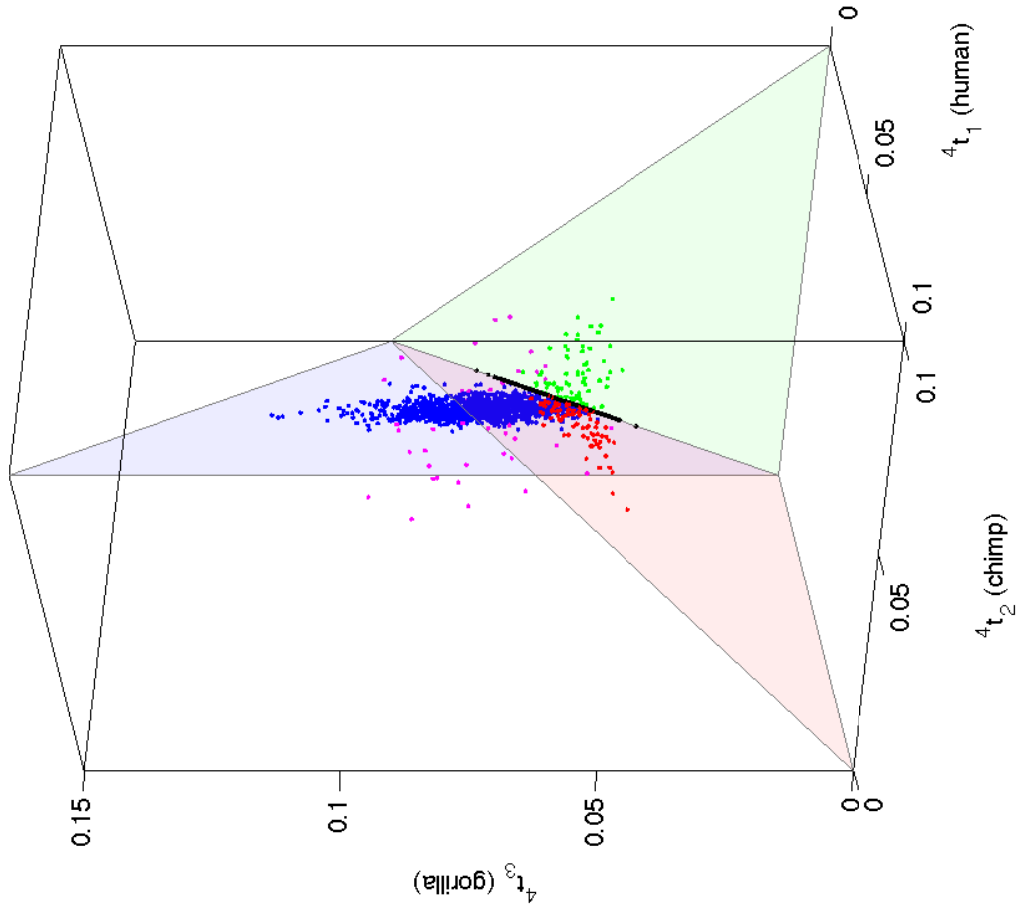


4MY H-C Div : $(272680, 571124, 1073375)$

8MY H-C Div : $(545360, 1142248, 2146749)$

MRS – Interval-extended Recursions on Trans-dimensional Tree Spaces





Model Selection in Primate Inter-relatedness across 3×10^9 Human-Chimp-Gorilla Genomes

$0.8668 \pm 0.0067, 0.1127 \pm 0.0062, 0.0086 \pm 0.0018, 0.0075 \pm 0.0017, 0.0044 \pm 0.0013$

Summary

- Advantages:
  - Exploits the DAG encoding of the function in the machine
  - Automatically constructs proposal density shape adapted to the target density
  - Produce guaranteed independent samples from inclusion isotonic densities
  - Can be used to account for physical limits on numerical and empirical resolutions in inferential procedures
- Limitations:
  - Can be overwhelmed by complexity, domain size, many dimensions.
  - Much work refining partition before first sample.
  - The MRS is ultimately RAM limited
  - Discrete spaces without an apparent metric structure ?
- POA:
  - Pre-enclosures – DAG dissection – Hash Access
  - Algebraic Statistics for dissolving symmetries in DAG (’minimal sufficiency’)
  - Tighter enclosures via AD – higher-order Taylor expansions
  - Extending arithmetic and $\mathfrak{E}$ to regular sub-pavings
- The support need not necessarily be Euclidean (CAT(0) space of trees is OK !)
- The puniest($< 1\ ulp$)-headed 11-dimensional witch "takes off her hat" to MRS !
- Moore Rejection Sampler is an Auto-validating von Neumann Rejection Sampler

http://www.math.canterbury.ac.nz/~r.sainudiin/codes/mrs/index.shtml

Acknowledgments

People:

Rick Durrett (MATH@Cornell) – *Listening and guiding*

Michael Nussbaum (MATH@Cornell) – *Math. stats. seminars & discussons*

Laurent Saloff-Coste (MATH@Cornell) – *MCMC asymptotics*

Rob Strawderman (STATISICS@Cornell) – *Finite mixtures*

Warwick Tucker (MATH@Uppsala, SW) – *Introduction to Interval analysis*

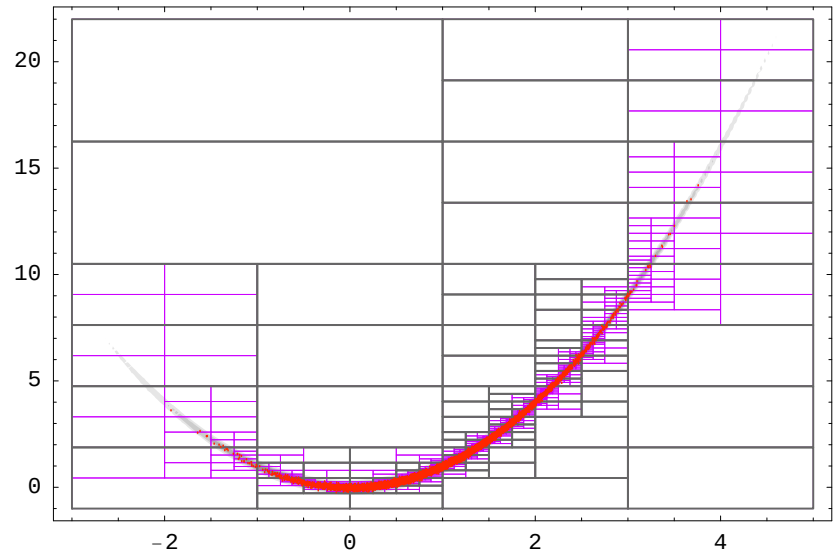
Karen Vogtmann (MATH@Cornell) – *Geometry of tree space*

Marty Wells (STATISTICS@Cornell) – *Various statistical insights*

Comp. Sci. Univ. of Karlsruhe, DL – > 25 *human years of prog. C++ librs.*

Funding:

Research Fellow of the Royal Commission for the Exhibition of 1851 (Oxford)



Just do it !–Sun-DARPA interval hardware in progress